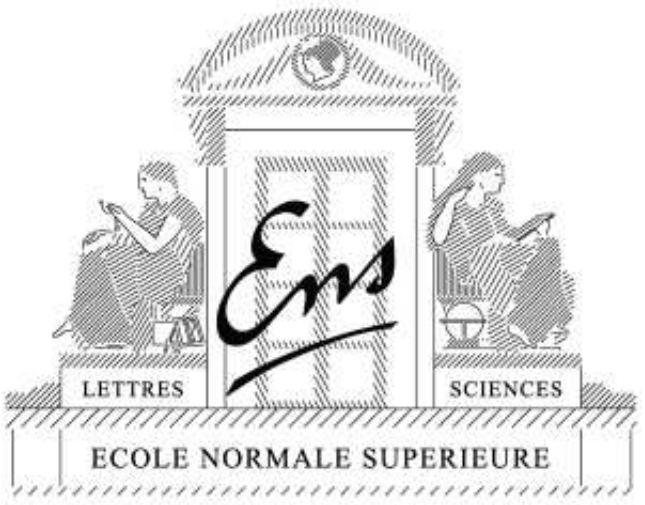
Structured sparse methods for matrix factorization

Francis Bach

Willow project, INRIA - Ecole Normale Supérieure



December 2010 - Joint work with R. Jenatton, J. Mairal, G. Obozinski, J. Ponce, G. Sapiro

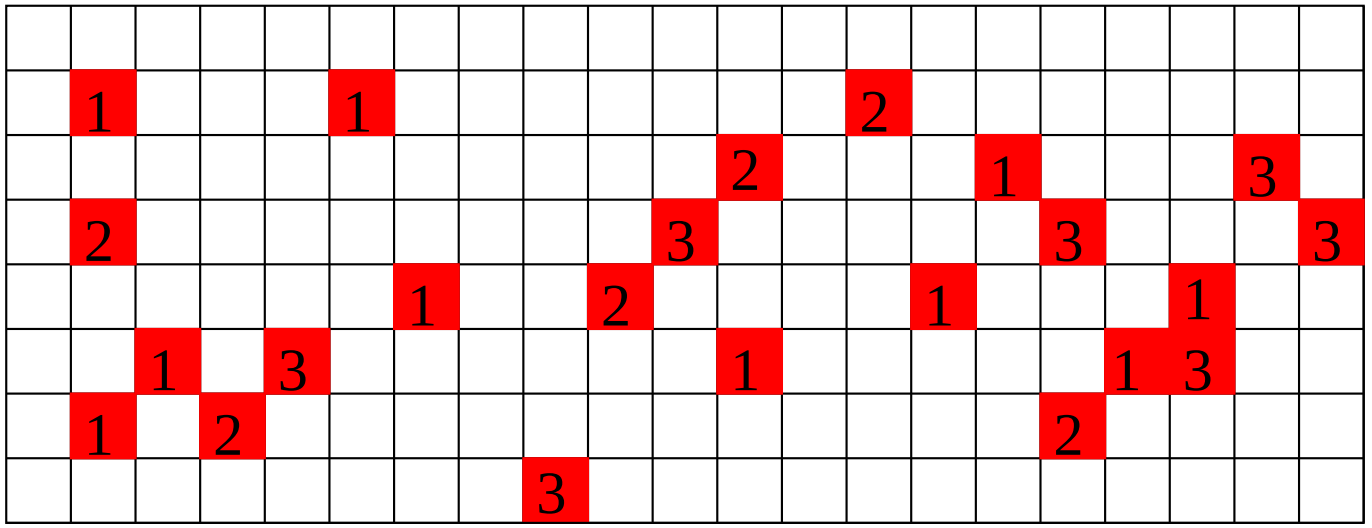
Structured sparse methods for matrix factorization

Outline

- **Learning problems on matrices**
- **Sparse methods for matrices**
  - Sparse principal component analysis
  - Dictionary learning
- **Structured sparse PCA/dictionary learning**
  - Structure on decomposition coefficients

Learning on matrices - Collaborative filtering

- Given $n_{\mathcal{X}}$ "movies" $\mathbf{x} \in \mathcal{X}$ and $n_{\mathcal{Y}}$ "customers" $\mathbf{y} \in \mathcal{Y}$,
- predict the "rating" $z(\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$ of customer $\mathbf{y}$ for movie $\mathbf{x}$
- Training data: large $n_{\mathcal{X}} \times n_{\mathcal{Y}}$ incomplete matrix $\mathbf{Z}$ that describes the known ratings of some customers for some movies
- **Goal**: complete the matrix.

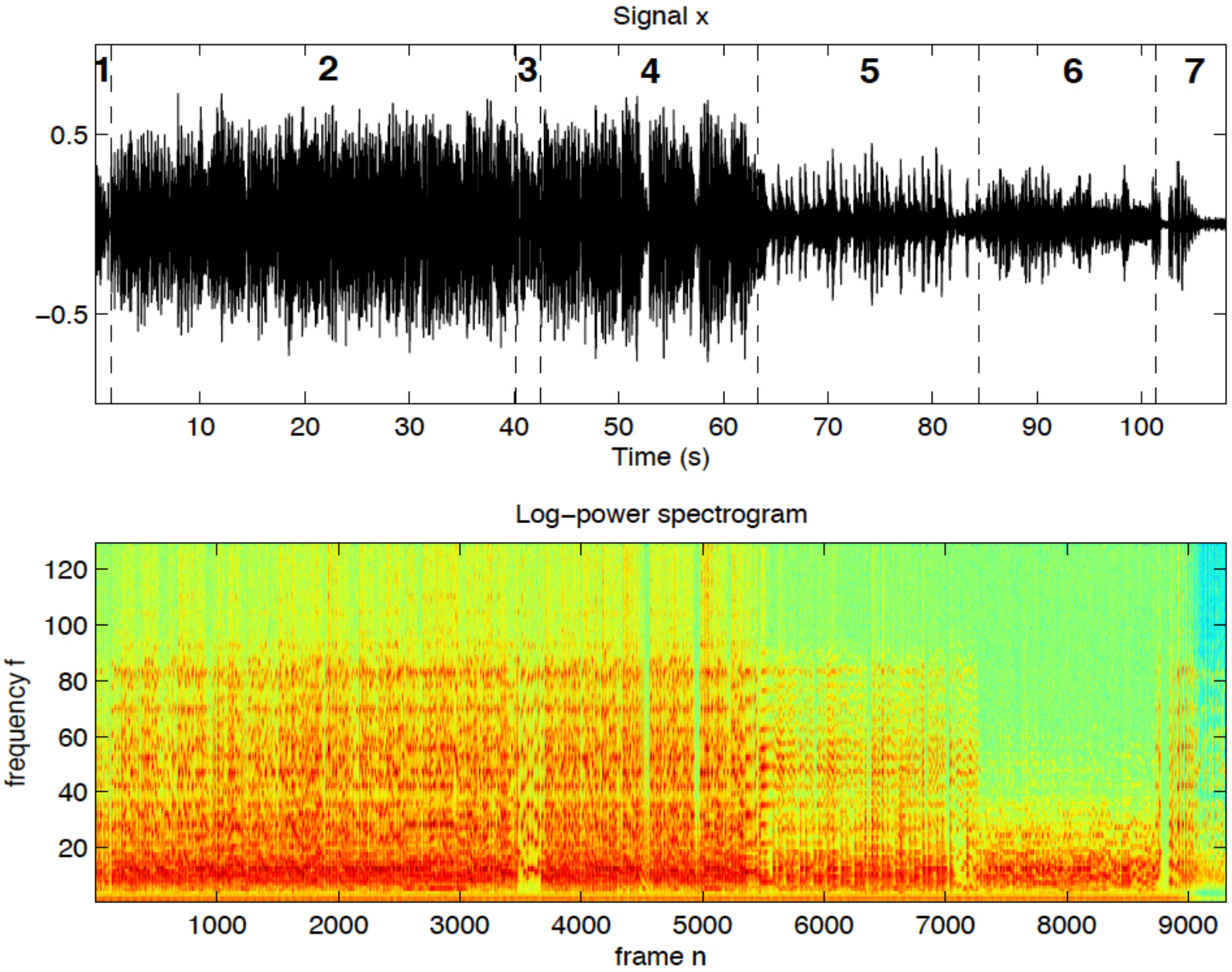


Learning on matrices - Image denoising

- Simultaneously denoise all patches of a given image
- Example from Mairal, Bach, Ponce, Sapiro, and Zisserman (2009b)

Learning on matrices - Source separation

- Single microphone (Benaroya et al., 2006; Févotte et al., 2009)

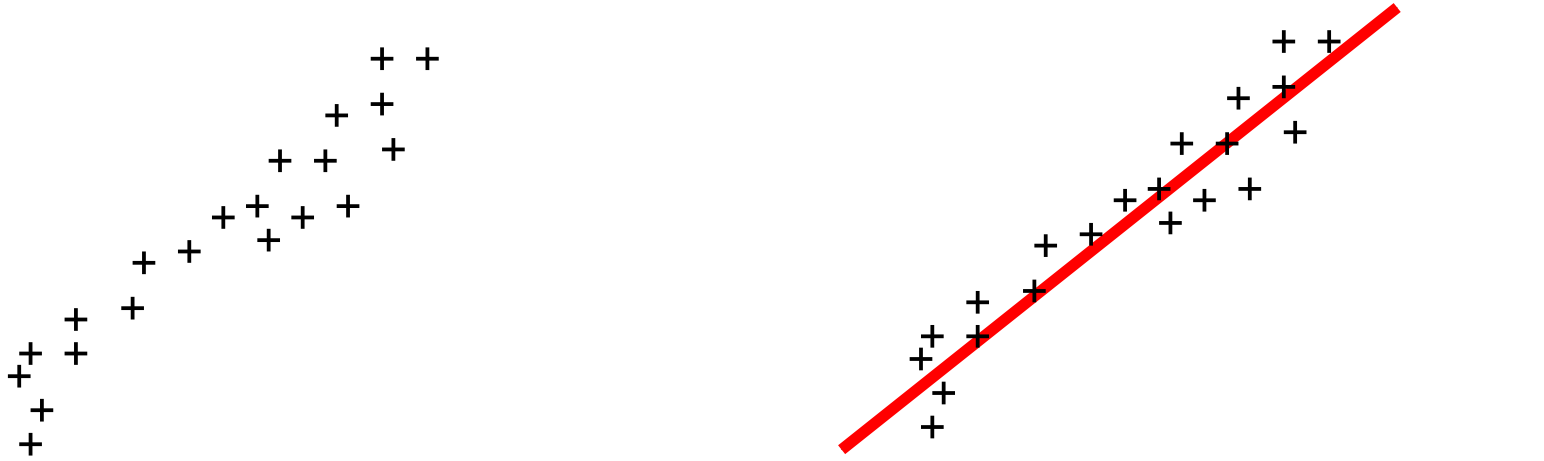


Learning on matrices - Multi-task learning

- k linear prediction tasks on same covariates $\mathbf{x} \in \mathbb{R}^p$
  - k weight vectors $\mathbf{w}_j \in \mathbb{R}^p$
  - Joint matrix of predictors $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{p \times k}$
- Classical applications
  - Transfer learning
  - Multi-category classification (one task per class) (Amit et al., 2007)
- **Share parameters between tasks**
  - Joint variable or feature selection (Obozinski et al., 2009; Pontil et al., 2007)

Learning on matrices - Dimension reduction

- Given data matrix $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top) \in \mathbb{R}^{n \times p}$
  - **Principal component analysis**: $\boxed{\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i}$

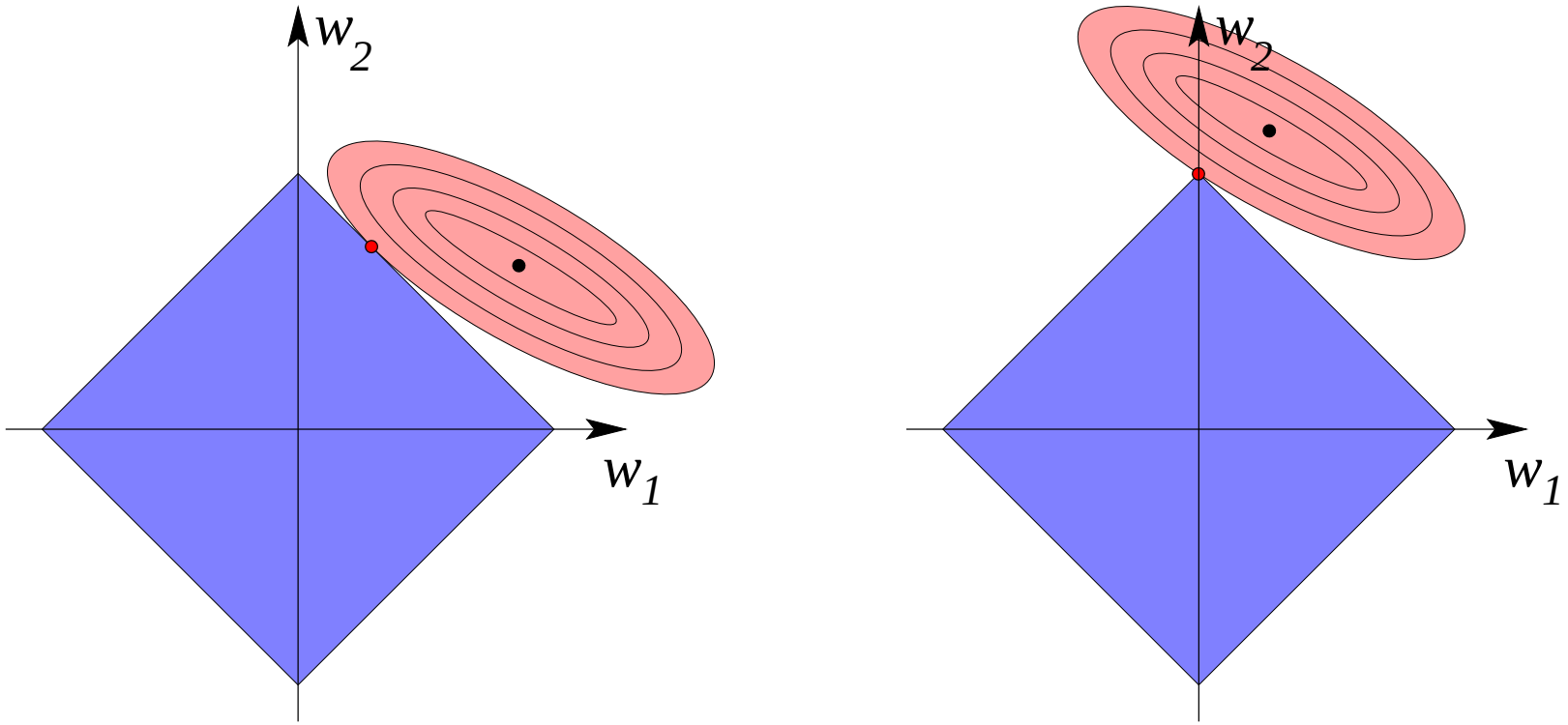


  - **K-means**: $\boxed{\mathbf{x}_i \approx \mathbf{d}_k \Rightarrow \mathbf{X} = \mathbf{DA}}$

Sparsity in machine learning

- **Assumption**: $\mathbf{y} = \mathbf{w}^\top \mathbf{x} + \varepsilon$, with $w \in \mathbb{R}^p$ sparse
  - Proxy for interpretability
  - Allow high-dimensional inference: $\boxed{\log p = O(n)}$

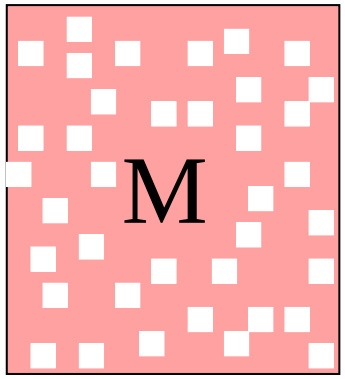
- **Sparsity and convexity** (ℓ_1-norm regularization): $\boxed{\min_{\mathbf{w} \in \mathbb{R}^p} L(\mathbf{w}) + \|\mathbf{w}\|_1}$



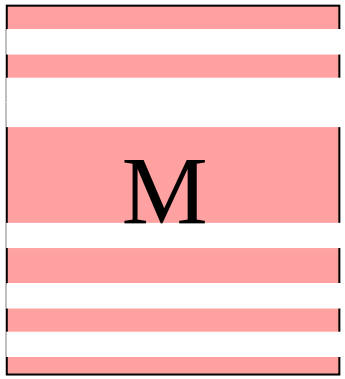
Two types of sparsity for matrices $\mathbf{M} \in \mathbb{R}^{n \times p}$

I - Directly on the elements of $\mathbf{M}$

- Many zero elements: $\mathbf{M}_{ij} = 0$

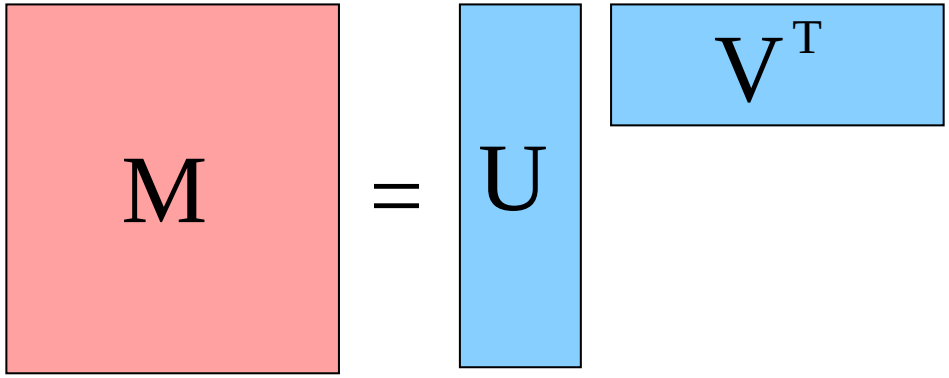


- Many zero rows (or columns): $(\mathbf{M}_{i1}, \ldots, \mathbf{M}_{ip}) = 0$

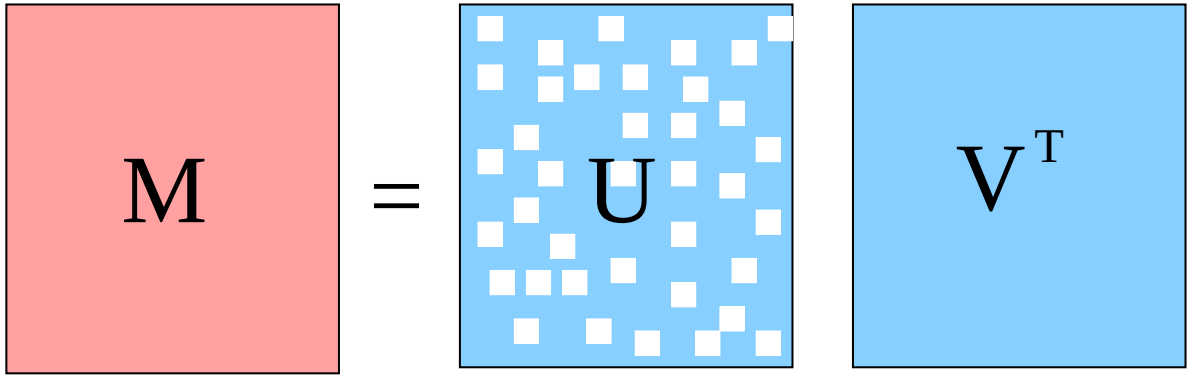


Two types of sparsity for matrices $\mathbf{M} \in \mathbb{R}^{n \times p}$
II - Through a factorization of $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$

- Matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times k}$ and $\mathbf{V} \in \mathbb{R}^{p \times k}$
- **Low rank**: m small



- **Sparse decomposition**: $\mathbf{U}$ sparse



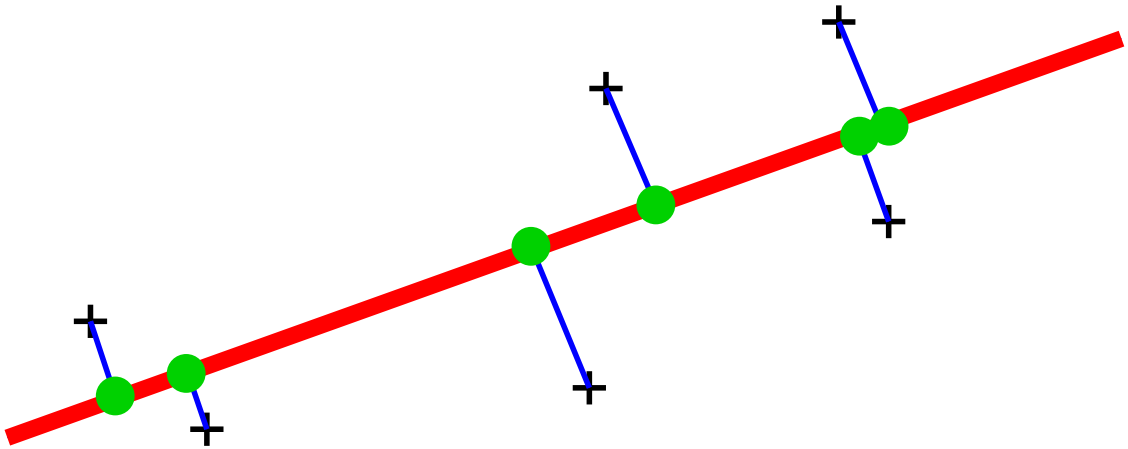
Structured sparse matrix factorizations

- Matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times k}$ and $\mathbf{V} \in \mathbb{R}^{p \times k}$
- **Structure on $\mathbf{U}$ and/or $\mathbf{V}$**
  - Low-rank: $\mathbf{U}$ and $\mathbf{V}$ have few columns
  - Dictionary learning / sparse PCA: $\mathbf{U}$ has many zeros
  - Clustering (k-means): $\mathbf{U} \in \{0,1\}^{n \times m}$, $\mathbf{U}\mathbf{1} = \mathbf{1}$
  - Pointwise positivity: non negative matrix factorization (NMF)
  - Specific patterns of zeros
  - Low-rank + sparse (Candès et al., 2009)
  - etc.
- **Many applications**
- **Many open questions**: Algorithms, identifiability, etc.

Sparse principal component analysis

- Given data $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top) \in \mathbb{R}^{p \times n}$, two views of PCA:
  - **Analysis view**: find the projection $\mathbf{d} \in \mathbb{R}^p$ of maximum variance (with deflation to obtain more components)
  - **Synthesis view**: find the basis $\mathbf{d}_1, \ldots, \mathbf{d}_k$ such that all $\mathbf{x}_i$ have low reconstruction error when decomposed on this basis

- For regular PCA, the two views are equivalent



Sparse principal component analysis

- Given data $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top) \in \mathbb{R}^{p \times n}$, two views of PCA:
  - **Analysis view**: find the projection $\mathbf{d} \in \mathbb{R}^p$ of maximum variance (with deflation to obtain more components)
  - **Synthesis view**: find the basis $\mathbf{d}_1, \ldots, \mathbf{d}_k$ such that all $\mathbf{x}_i$ have low reconstruction error when decomposed on this basis
- For regular PCA, the two views are equivalent
- **Sparse extensions**
  - Interpretability
  - High-dimensional inference
  - Two views are differents
    * For analysis view, see d'Aspremont, Bach, and El Ghaoui (2008); Journée, Nesterov, Richtárik, and Sepulchre (2010)

Sparse principal component analysis
Synthesis view

- Find $\mathbf{d}_1, \ldots, \mathbf{d}_k \in \mathbb{R}^p$ **sparse** so that

$$\sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \sum_{j=1}^{k} (\boldsymbol{\alpha}_i)_j \mathbf{d}_j \right\|_2^2 = \sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i \right\|_2^2 \text{ is small}$$

  - Look for $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n) \in \mathbb{R}^{k \times n}$ and $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_k) \in \mathbb{R}^{p \times k}$ such that $\mathbf{D}$ is sparse and $\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2$ is small

Sparse principal component analysis
Synthesis view

- Find $\mathbf{d}_1, \ldots, \mathbf{d}_k \in \mathbb{R}^p$ **sparse** so that

$$\sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \sum_{j=1}^{k} (\boldsymbol{\alpha}_i)_j \mathbf{d}_j \right\|_2^2 = \sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 \text{ is small}$$

  - Look for $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n) \in \mathbb{R}^{k \times n}$ and $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_k) \in \mathbb{R}^{p \times k}$ such that $\mathbf{D}$ is sparse and $\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2$ is small

- Sparse formulation (Witten et al., 2009; Bach et al., 2008)
  - Penalize/constrain $\mathbf{d}_j$ by the ℓ_1-norm for sparsity
  - Penalize/constrain $\boldsymbol{\alpha}_i$ by the ℓ_2-norm to avoid trivial solutions

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_1 \text{ s.t. } \forall i, \|\boldsymbol{\alpha}_i\|_2 \leqslant 1$$

Structured matrix factorizations (Bach et al., 2008)

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_\star \text{ s.t. } \forall i, \|\boldsymbol{\alpha}_i\|_\bullet \leqslant 1$$

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_\bullet \text{ s.t. } \forall j, \|\mathbf{d}_j\|_\star \leqslant 1$$

- Optimization by alternating minimization (non-convex)
- $\boldsymbol{\alpha}_i$ decomposition coefficients (or "code"), $\mathbf{d}_j$ dictionary elements
- Two related/equivalent problems:
  - **Sparse PCA = sparse dictionary** (ℓ_1-norm on $\mathbf{d}_j$)
  - **Dictionary learning = sparse decompositions** (ℓ_1-norm on $\boldsymbol{\alpha}_i$) (Olshausen and Field, 1997; Elad and Aharon, 2006; Lee et al., 2007)

Dictionary learning for image denoising

$$\underbrace{\mathbf{x}}_{\text{measurements}} = \underbrace{\mathbf{y}}_{\text{original image}} + \underbrace{\varepsilon}_{\text{noise}}$$

Dictionary learning for image denoising

- **Solving the denoising problem** (Elad and Aharon, 2006)
  - Extract all overlapping 8×8 patches $\mathbf{x}_i \in \mathbb{R}^{64}$
  - Form the matrix $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top) \in \mathbb{R}^{n \times 64}$
  - Solve a matrix factorization problem:

$$\min_{\mathbf{D},\mathbf{A}} ||\mathbf{X} - \mathbf{D}\mathbf{A}||_F^2 = \min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2$$

    where $\mathbf{A}$ is **sparse**, and $\mathbf{D}$ is the **dictionary**
  - Each patch is decomposed into $\mathbf{x}_i = \mathbf{D}\boldsymbol{\alpha}_i$
  - Average the reconstruction $\mathbf{D}\boldsymbol{\alpha}_i$ of each patch $\mathbf{x}_i$ to reconstruct a full-sized image

- The number of patches n is large ($=$ number of pixels)

Online optimization for dictionary learning

$$\min_{\mathbf{A}\in\mathbb{R}^{k\times n},\mathbf{D}\in\mathcal{D}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda||\boldsymbol{\alpha}_i||_1$$

$$\mathcal{D} \triangleq \{\mathbf{D} \in \mathbb{R}^{p\times k} \text{ s.t. } \forall j = 1,\ldots,k, \quad ||\mathbf{d}_j||_2 \leqslant 1\}.$$

- Classical optimization alternates between $\mathbf{D}$ and $\mathbf{A}$
- Good results, but very slow !

Online optimization for dictionary learning

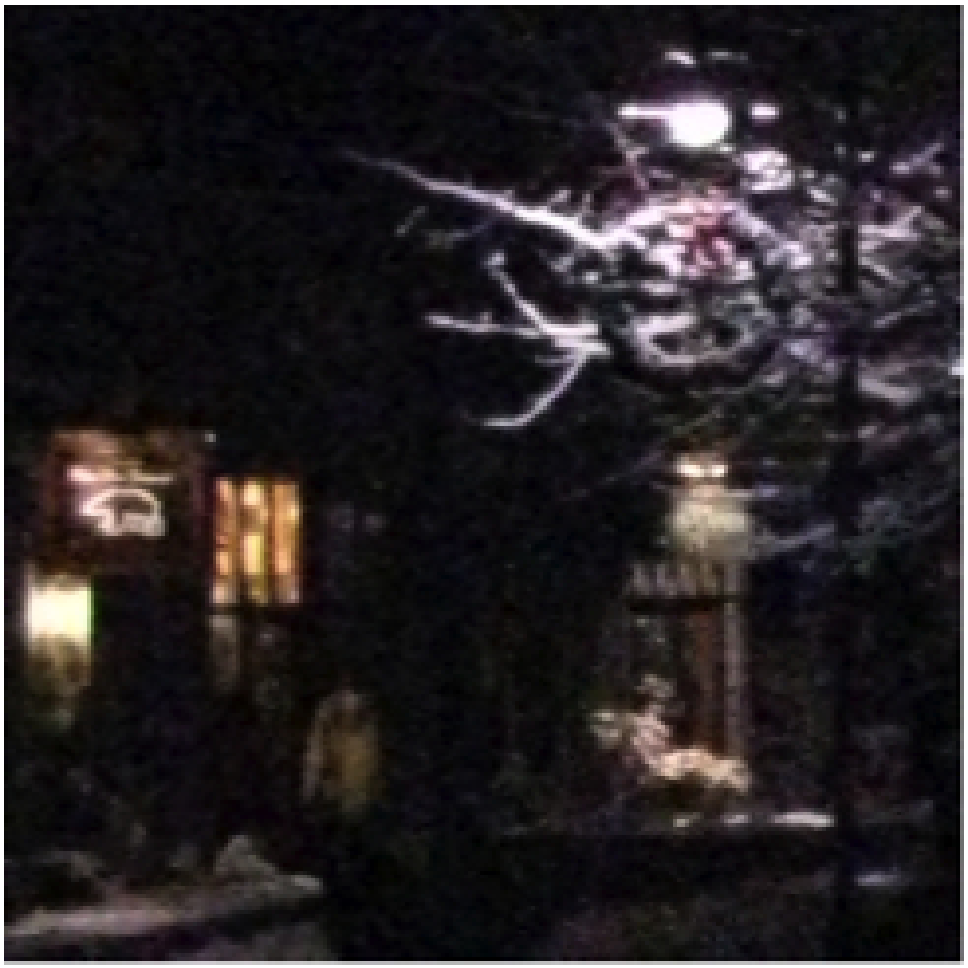
$$\min_{\mathbf{A}\in\mathbb{R}^{k\times n},\mathbf{D}\in\mathcal{D}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda||\boldsymbol{\alpha}_i||_1$$

$$\mathcal{D} \triangleq \{\mathbf{D} \in \mathbb{R}^{p\times k} \text{ s.t. } \forall j = 1,\ldots,k, \quad ||\mathbf{d}_j||_2 \leqslant 1\}.$$

- Classical optimization alternates between $\mathbf{D}$ and $\mathbf{A}$.
- Good results, but very slow !
- **Online learning** (Mairal, Bach, Ponce, and Sapiro, 2009a) can
  - handle potentially infinite datasets
  - adapt to dynamic training sets

Denoising result
(Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009b)

Denoising result
(Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009b)



Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

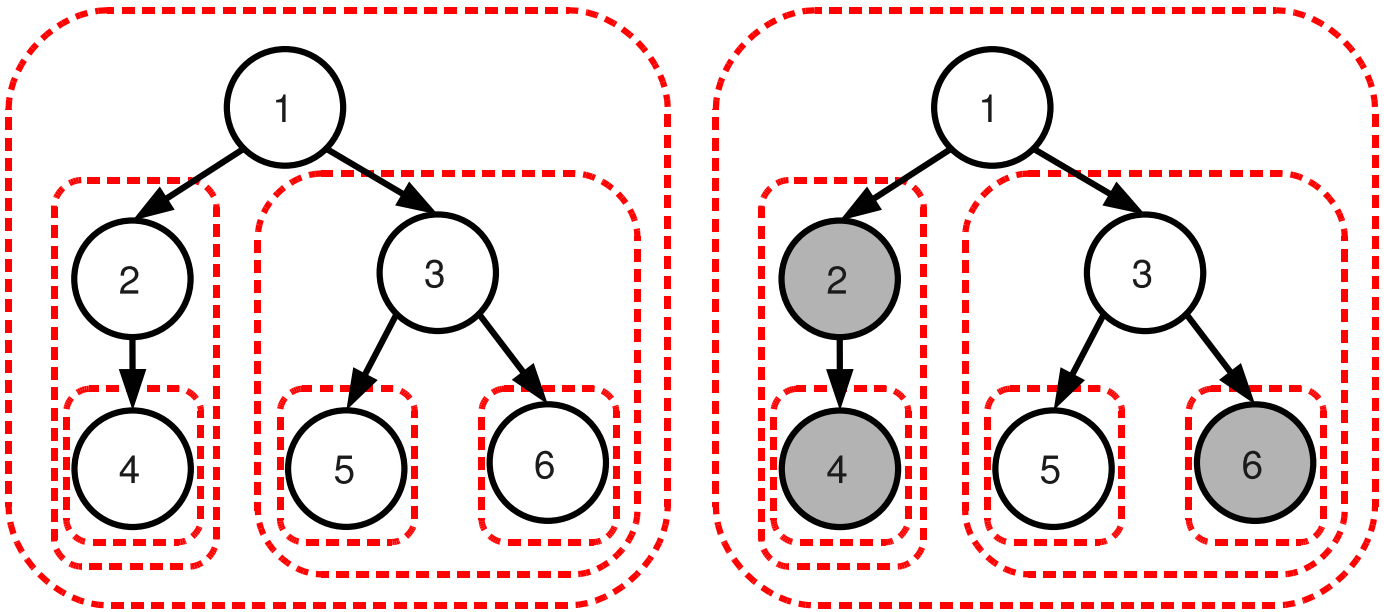
Structured sparse methods for matrix factorization

Outline

- **Learning problems on matrices**
- **Sparse methods for matrices**
  - Sparse principal component analysis
  - Dictionary learning
- **Structured sparse PCA/dictionary learning**
  - Structure on decomposition coefficients

Hierarchical dictionary learning (Jenatton, Mairal, Obozinski, and Bach, 2010)

- Structure on codes $\boldsymbol{\alpha}$ (not on dictionary $\mathbf{D}$)
- Hierarchical penalization: $\psi(\boldsymbol{\alpha}) = \sum_{G \in \mathbf{G}} \|\boldsymbol{\alpha}_G\|_2$ where groups G in $\mathbf{G}$ are equal to set of descendants of some nodes in a tree



- Variable selected after its ancestors (Zhao et al., 2009; Bach, 2008)

Hierarchical dictionary learning
Efficient optimization

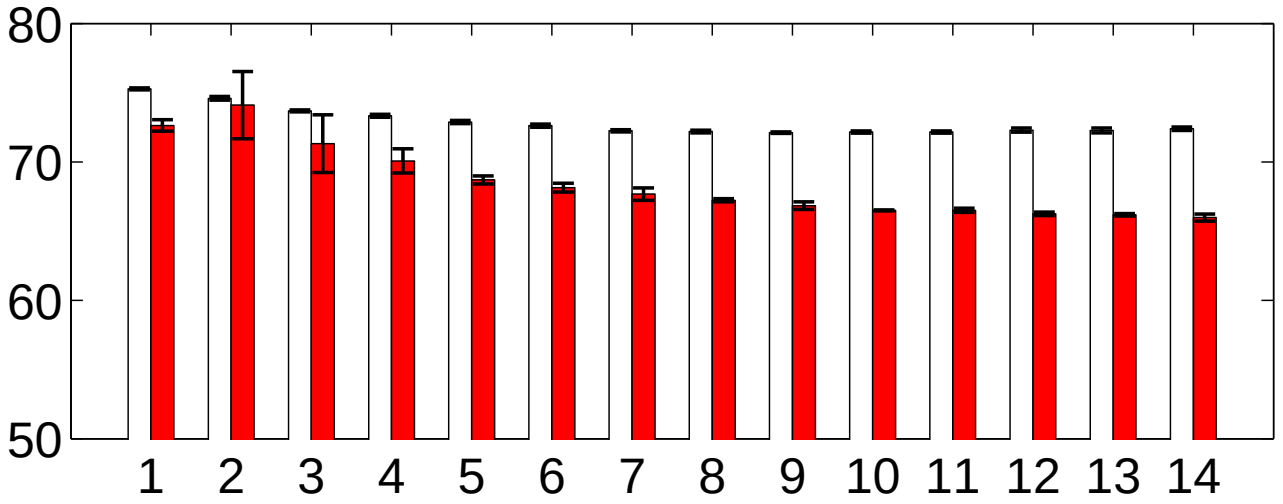
$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{k\times n}\\ \mathbf{D}\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\psi(\boldsymbol{\alpha}_i) \text{ s.t. } \forall j,\ \|\mathbf{d}_j\|_2 \leq 1.$$

- Minimization with respect to $\boldsymbol{\alpha}_i$: regularized least-squares
  - Many algorithms dedicated to the ℓ_1-norm $\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$
- Proximal methods : first-order methods with optimal convergence rate (Nesterov, 2007; Beck and Teboulle, 2009)
  - Requires solving many times $\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \boldsymbol{\alpha}\|_2^2 + \lambda\psi(\boldsymbol{\alpha})$
- Tree-structured regularization : **Efficient linear time algorithm based on primal-dual decomposition** (Jenatton et al., 2010)

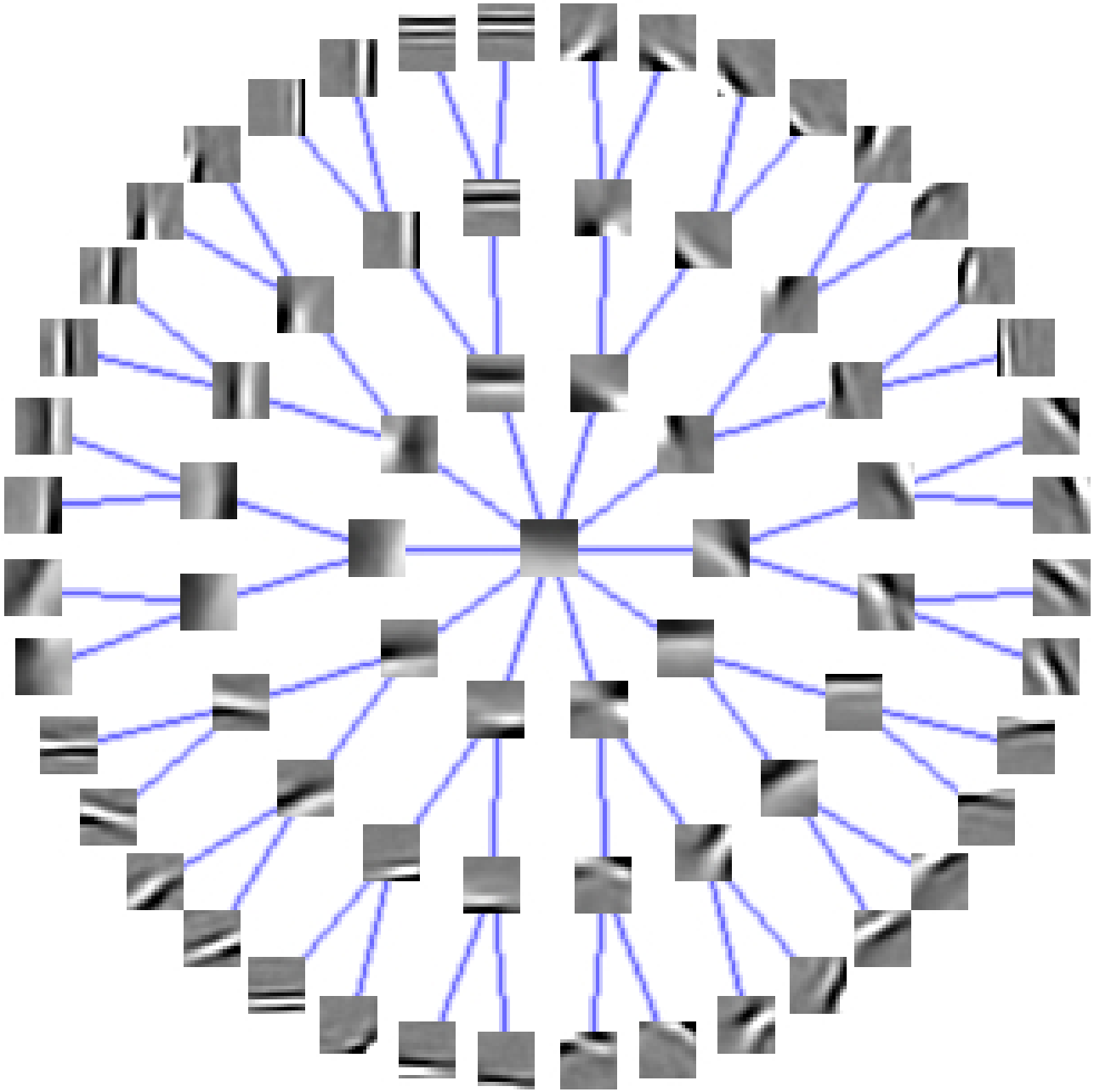
Hierarchical dictionary learning
Application to image denoising

- Reconstruction of 100,000 8×8 natural images patches
  - Remove randomly subsampled pixels
  - Reconstruct with matrix factorization and structured sparsity

noise	50 %	60 %	70 %	80 %	90 %
flat	19.3 ± 0.1	26.8 ± 0.1	36.7 ± 0.1	50.6 ± 0.0	72.1 ± 0.0
tree	18.6 ± 0.1	25.7 ± 0.1	35.0 ± 0.1	48.0 ± 0.0	65.9 ± 0.3



Application to image denoising - Dictionary tree



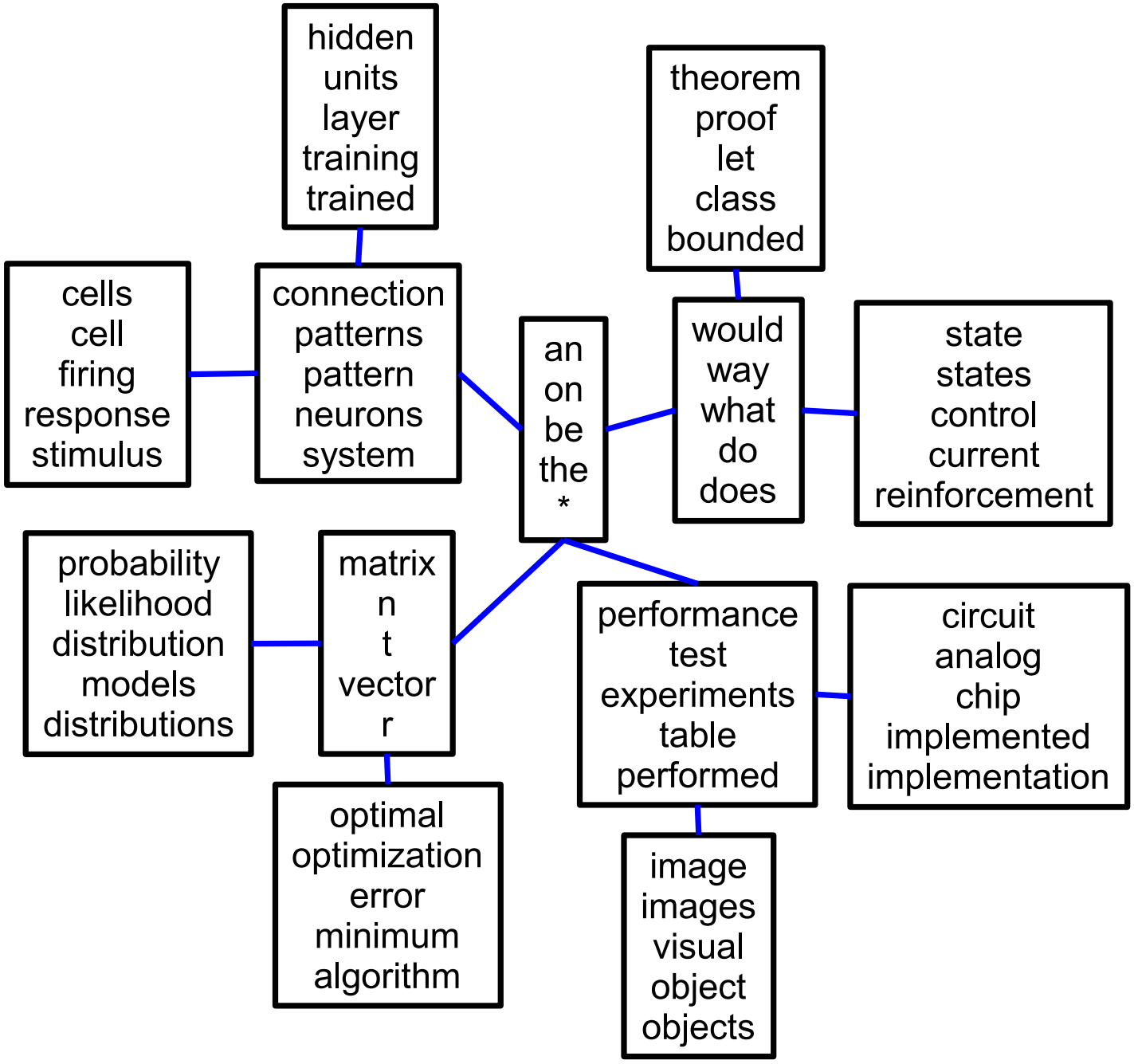
Hierarchical dictionary learning
Modelling of text corpora

- Each document is modelled through word counts
- Low-rank matrix factorization of word-document matrix
- Probabilistic topic models (Blei et al., 2003)
  - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
  - **Can we achieve similar performance with simple matrix factorization formulation?**

Hierarchical dictionary learning
Modelling of text corpora

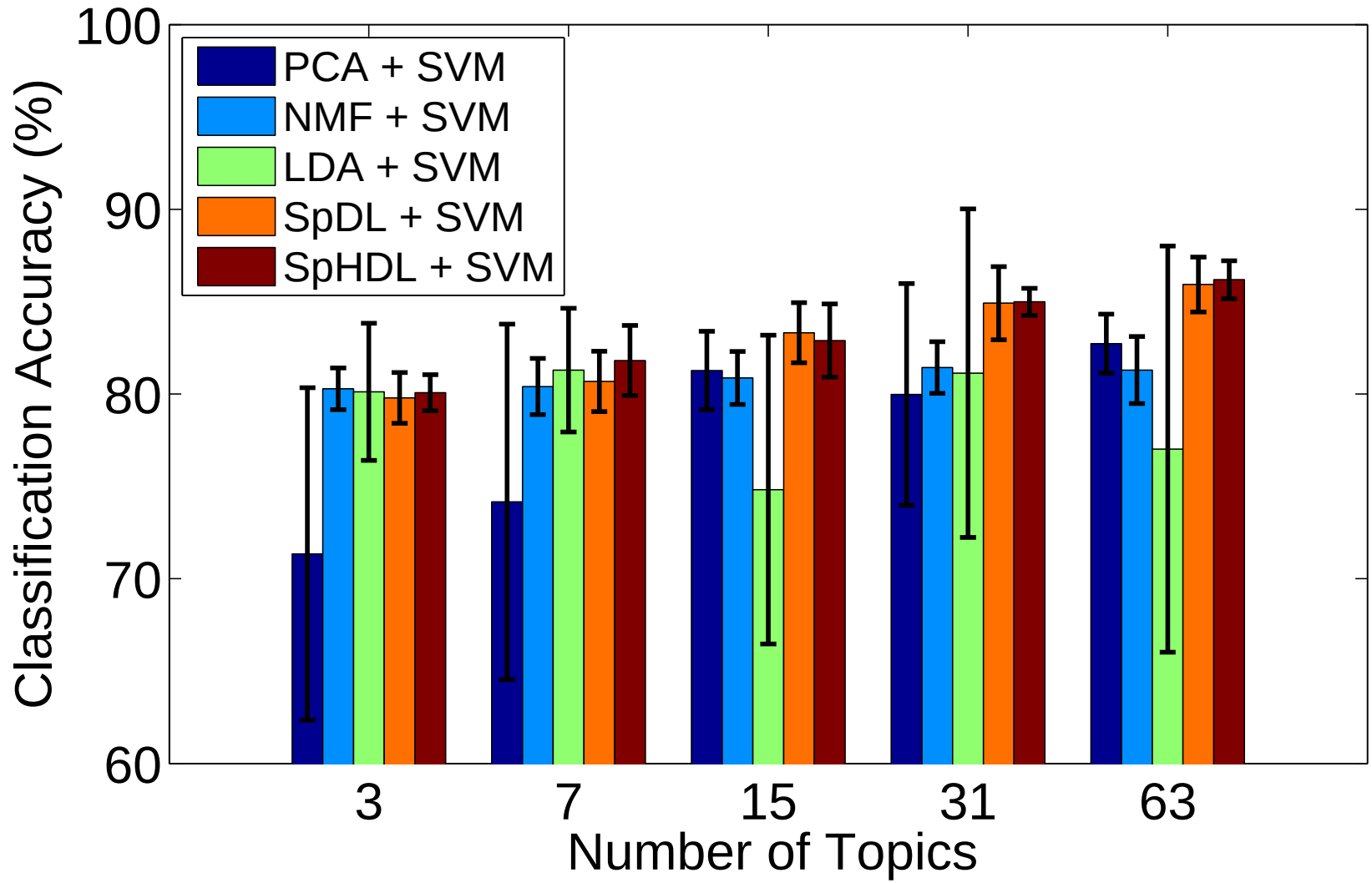
- Each document is modelled through word counts
- Low-rank matrix factorization of word-document matrix
- Probabilistic topic models (Blei et al., 2003)
  - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
  - **Can we achieve similar performance with simple matrix factorization formulation?**
- Experiments:
  - Qualitative: NIPS abstracts (1714 documents, 8274 words)
  - Quantitative: newsgroup articles (1425 documents, 13312 words)

Modelling of text corpora - Dictionary tree



Modelling of text corpora

- Comparison on predicting newsgroup article subjects:



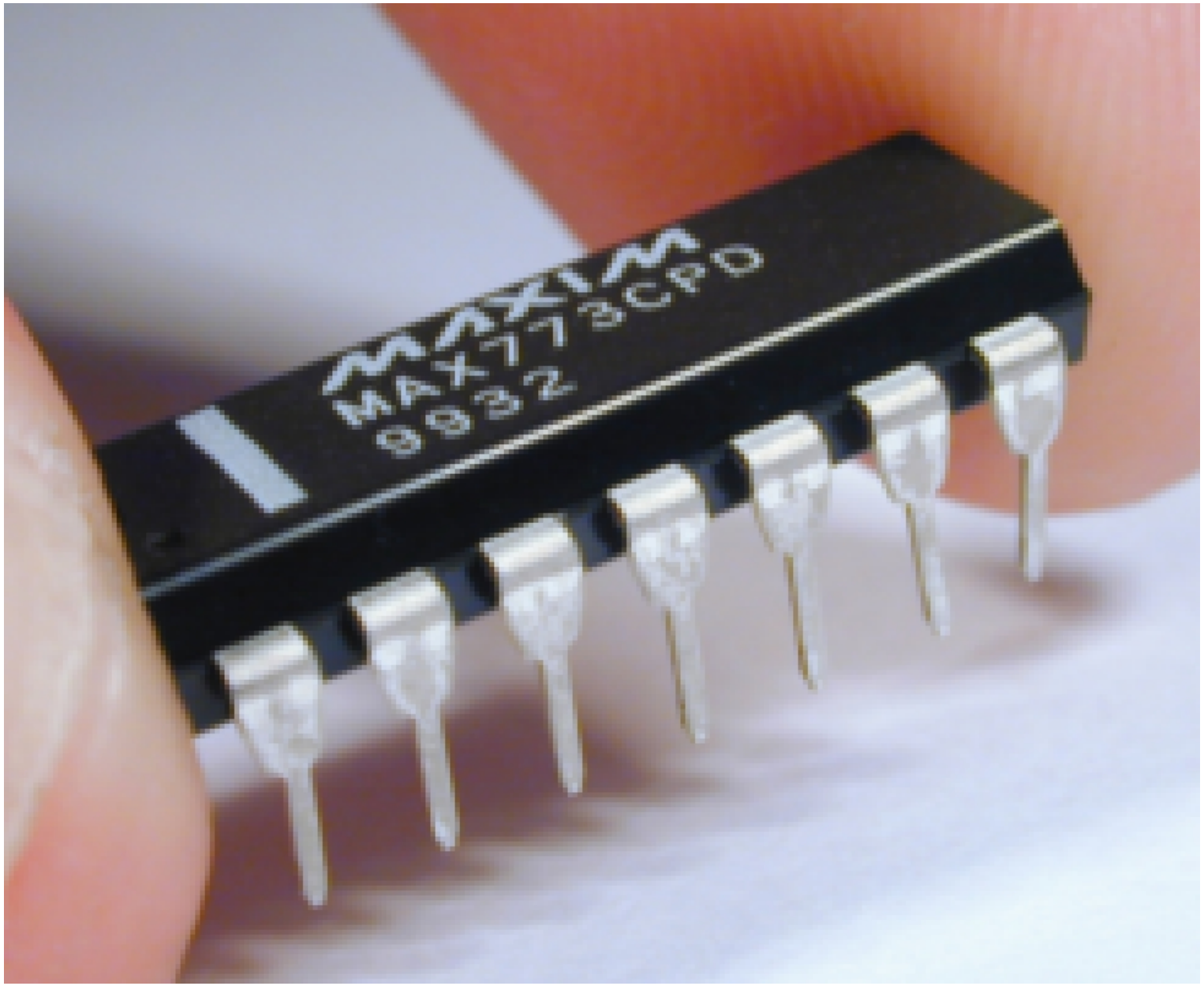
Conclusion

- Structured matrix factorization has many applications
  - Machine learning
  - Image/signal processing
  - Extensions to other tasks
- Algorithmic issues
  - Large datasets
  - Structured sparsity and convex optimization
- Theoretical issues
  - Identifiability of structures and features
  - Improved predictive performance
  - Other approaches to sparsity and structure (e.g., submodularity)

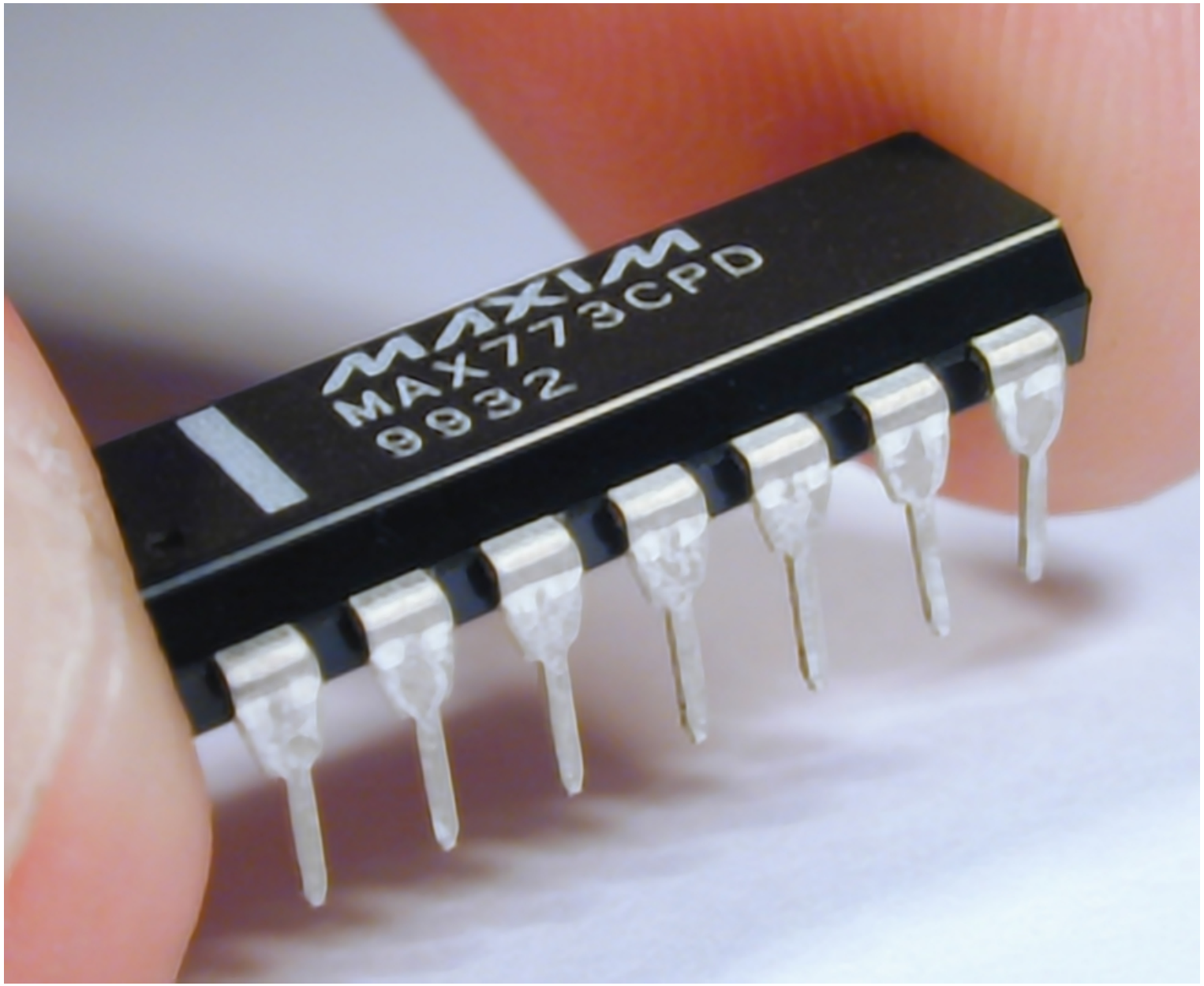
Ongoing Work - Digital Zooming

Digital Zooming (Couzinie-Devy et al., 2010)

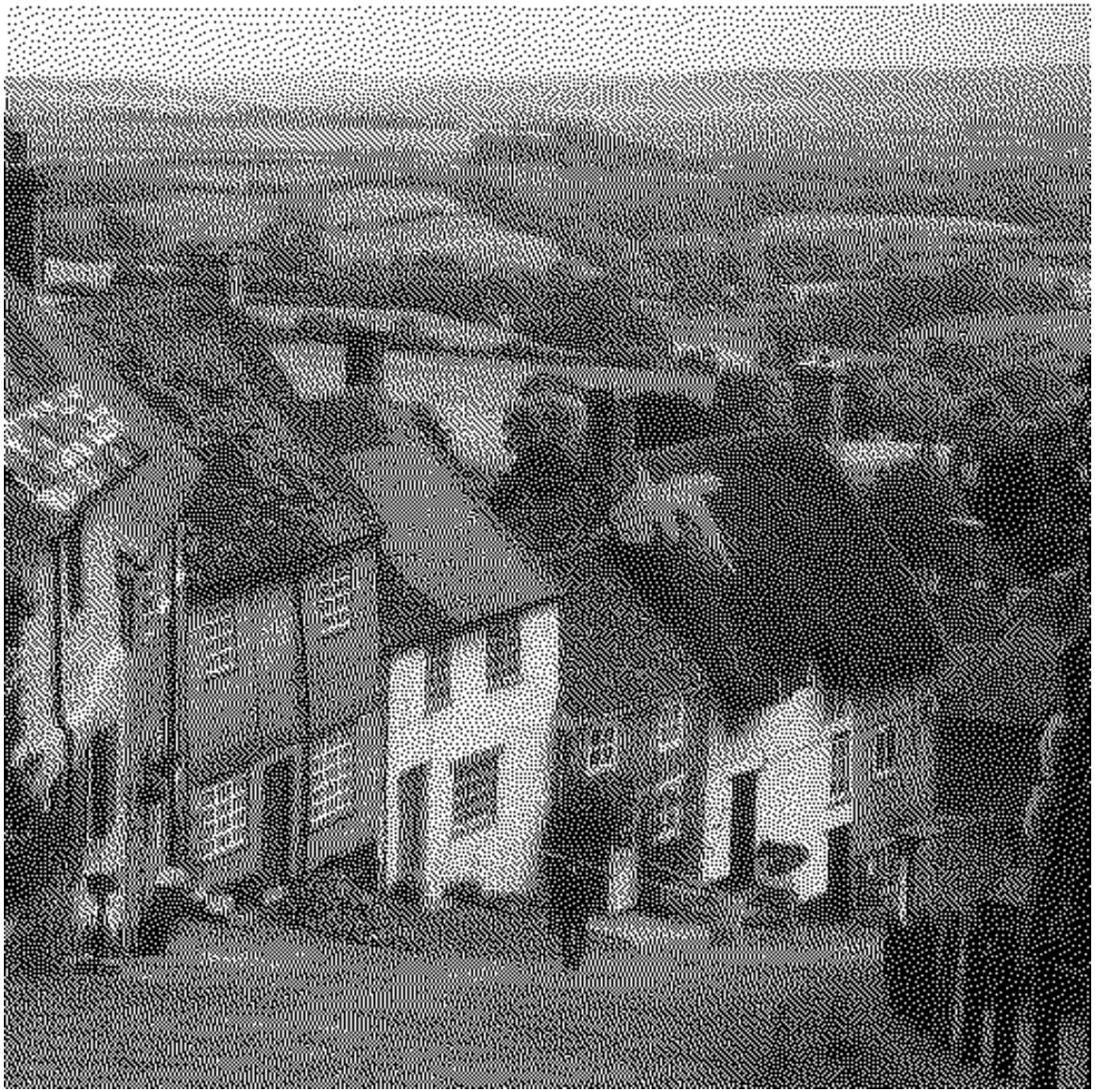
Digital Zooming (Couzinie-Devy et al., 2010)



Digital Zooming (Couzinie-Devy et al., 2010)

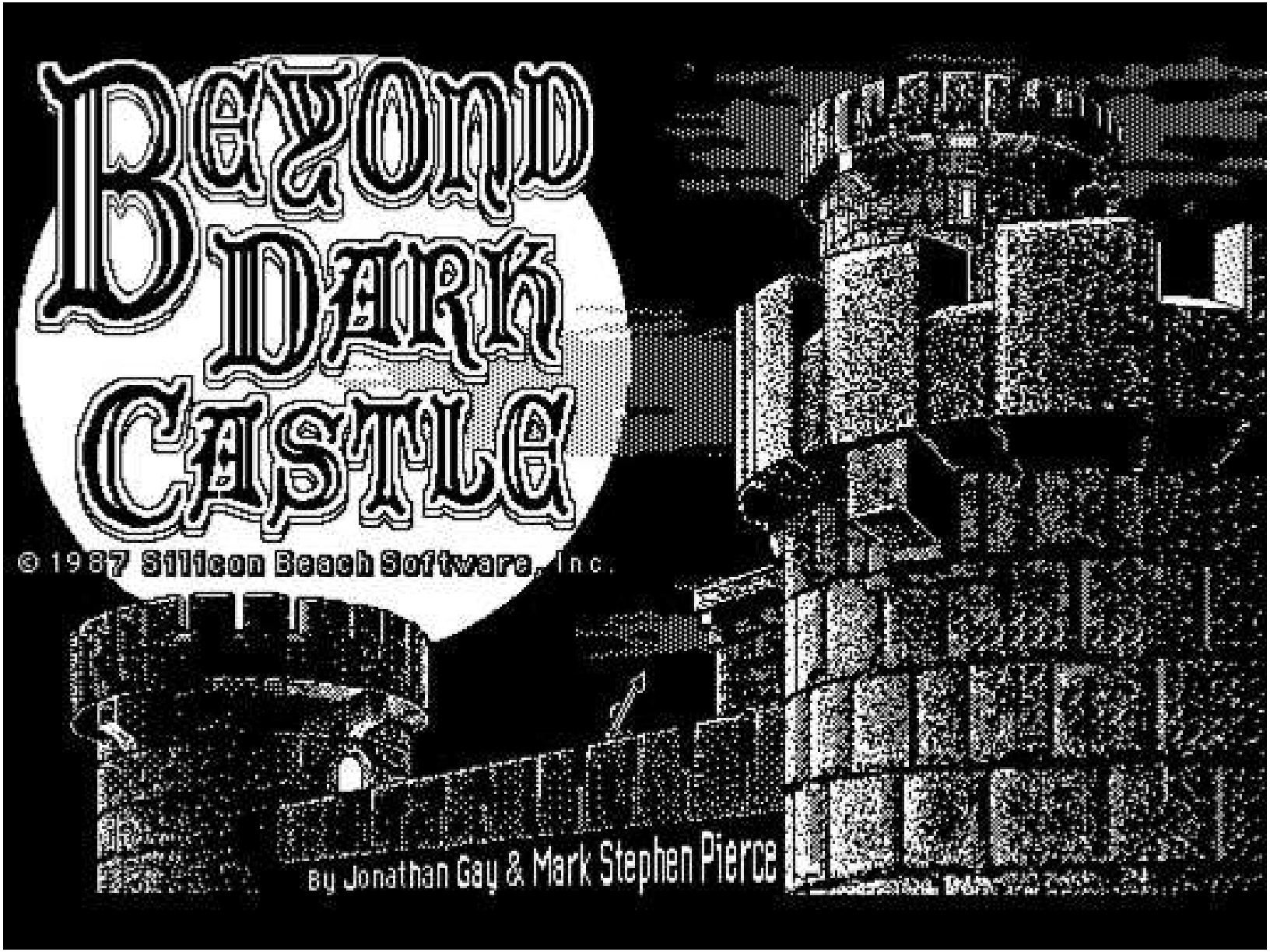


Ongoing Work - Task-driven dictionaries inverse half-toning (Mairal et al., 2010)



Ongoing Work - Task-driven dictionaries inverse half-toning (Mairal et al., 2010)

Ongoing Work - Inverse half-toning



Ongoing Work - Inverse half-toning

Ongoing Work - Inverse half-toning

Ongoing Work - Inverse half-toning

References

Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine Learning (ICML)*, 2007.

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008.

F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. Technical Report 0812.1869, ArXiv, 2008.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Speech and Audio Processing*, 14(1):191, 2006.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, January 2003.

D. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16:106, 2004.

A. Buades, B. Coll, and J.-M. Morel. Non-local image and movie denoising. *International Journal of Computer vision*, 76(2):123–139, 2008.

E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Arxiv preprint arXiv:0912.3599*, 2009.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis. *Neural Computation*, 21(3), 2009.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Submitted to ICML*, 2010.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized Power Method for Sparse Principal Component Analysis. *Journal of Machine Learning Research*, 11:517–553, 2010.

H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *International Conference on Machine Learning (ICML)*, 2009a.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *International Conference on Computer Vision (ICCV)*, 2009b.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.

G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 1–22, 2009.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed

by V1? *Vision Research*, 37:3311–3325, 1997.

M. Pontil, A. Argyriou, and T. Evgeniou. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, 2007.

D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.